The evaluation of evidence for exponentially distributed data

Shen, Qiang; Hayes, B.; Aitken, Colin; Jensen, Richard

Published in:
Computational Statistics and Data Analysis

DOI:
10.1016/j.csda.2007.05.026

Publication date:
2007

Citation for published version (APA):
Shen, Q., Hayes, B., Aitken, C., & Jensen, R. (2007). The evaluation of evidence for exponentially distributed data. *Computational Statistics and Data Analysis*, *51*(12), 5682-5693. https://doi.org/10.1016/j.csda.2007.05.026

The evaluation of evidence for exponentially distributed data

C.G.G. Aitken,
School of Mathematics, King's Buildings,
Mayfield Road, Edinburgh, EH9 3JZ;
c.g.g.aitken@ed.ac.uk
Q. Shen,
Department of Computer Science,
The University of Wales, Aberystwyth;
qqs@aber.ac.uk
R. Jensen,
Department of Computer Science,
The University of Wales, Aberystwyth;
rkj@aber.ac.uk
B. Hayes,
School of Mathematics, King's Buildings,
Mayfield Road, Edinburgh, EH9 3JZ.

January 8, 2007

Abstract

At present, likelihood ratios for two-level models are determined with the use of a normal kernel estimation procedure when the between-group distribution is thought to be non-normal. An extension is described here for a two-level model in which the between-group distribution is very positively skewed and an exponential distribution may be thought to represent a good model. The theoretical likelihood ratio is derived. A likelihood ratio based on a biweight kernel with an adaptation at the boundary is developed. The performance of this kernel is compared alongside those of normal kernels and normal and exponential parametric models. A comparison of performance is made for simulated data where results may be compared with those of theory, using the theoretical model, as the true parameter values for the models are known. There is also a comparison for forensic data, using the concentration of aluminium in glass as an exemplar. Performance is assessed by determining the numbers of occasions on which the likelihood ratios for sets of fragments from the same group are supportive of the proposition that they are from different groups and the

numbers of occasions on which the likelihood ratios for sets of fragments from different group are supportive of the proposition that they are from the same group.

1 Introduction

The value of evidence, E, in comparing the probabilities of the truth of two propositions, H_p and H_d say, is taken to be the factor which converts the odds in favour of H_p, relative to H_d, prior to consideration of E, to the odds in favour of H_p, relative to H_d, posterior to consideration of E. From the odds form of Bayes' Theorem, the value of the evidence can be seen to be the likelihood ratio

$$\frac{Pr(E \mid H_p)}{Pr(E \mid H_d)}$$

such that

$$\frac{Pr(H_p \mid E)}{Pr(H_d \mid E)} \frac{Pr(E \mid H_p)}{Pr(E \mid H_d)} \times \frac{Pr(H_p)}{Pr(H_d)}. \tag{1}$$

Trace evidence, as the name suggests, is evidence which is found in traces, for example, stains of body fluids such as blood, or fragments of glass or a pile of powdered drugs. Evidence whose source is known, such as fragments of glass taken from a window at a crime scene is known as control evidence. Evidence whose source is unknown, such as fragments of glass taken from the clothing of a person suspected of committing the crime is known as recovered evidence. Some evidence is in the form of measurements, such as the elemental composition of glass or the chemical composition of drugs. The data from these measurements are often nested with two levels. There are measurements from within a source, such as from a single window, and measurements between

sources, such as between different windows. Methods have been developed for the evaluation of evidence where the data are univariate and the within- and between-group distributions are both normal (Lindley, 1977) and where the data are multivariate, the within-group distribution is normal and the between-group distribution is non-normal (Aitken and Taroni, 2004, Aitken *et al.*, 2007). When the between-group distribution is non-normal, the distribution has been estimated by a Gaussian kernel function.

The method described here for the evaluation of evidence is applicable to a univariate two-level model where the within-group distribution is taken to be normal and the between-group distribution is very highly positively skewed. It is not amenable to a simple transformation to normality nor can it be modelled satisfactorily by a Gaussian kernel function. An example is given of the between-group distribution of the concentration of aluminium in glass which is very positively skewed (see Figure 1). A closed-form expression is derived here for the value of the evidence when the between-group distribution is exponential. A kernel estimator, for incorporation in the expression for the likelihood ratio, is developed based on biweight and boundary kernels (Silverman, 1986).

Simulation studies are carried out to compare various models for the value of the evidence in different scenarios, based on different estimates for the between-group distribution. Experimental results show within-group distributions to be normal. The scenarios relate to the similarity to each other of the control and recovered data and their rarity. Control and recovered data which are similar and rare are expected to have a high value, namely a likelihood ratio consider-

ably greater than one in (1). Control and recovered data which are dissimilar are expected to have a low value, namely a likelihood ratio considerably less than one in (1). Models compared with the theoretical results estimate the between-group distribution with (a) a normal distribution, (b) a normal kernel function, (c) an adaptive kernel function with several choices of sensitivity parameter and (d) a biweight kernel. The methods are then applied to data of elemental concentrations of aluminium.

The rest of the paper is developed as follows. Section 2 gives the derivation of the likelihood ratio in (1) in an analytical form when the between-group distribution is exponential. A method for the estimation of the likelihood ratio for highly skewed data is given in Section 3 using biweight and boundary kernels. In Section 4, the performances of various methods of estimating the likelihood ratio are assessed using simulations of various combinations of control and recovered data which compare similarity and rarity. Section 5 provides an assessments of the performances of the various methods using the example of the concentration of aluminium in glass, as illustrated in Figure 1. Some conclusions are given in Section 6 and an Appendix gives a few lines to explain the derivation of the variance of the biweight kernel.

2 Derivation of likelihood ratio

Consider a two-level random effects model for a random variable X such that $(X_{ij} \mid \mu_i, \sigma^2)$ is normally distributed with expectation μ_i and variance σ^2 and μ_i is exponentially distributed with expectation α^{-1} with probability density

function

$$f(\mu \mid \alpha) = \alpha \exp(-\alpha\mu).$$

The variance of μ_i is $1/\alpha^2$.

Let $\{x_{ij}, i = 1, \ldots, m,\ j = 1, \ldots, k\}$ be a random sample from this model of k observations from each of m groups. Denote the m group means by $\bar{x}_1, \ldots, \bar{x}_m$ where $\bar{x}_i = \sum_{j=1}^{k} x_{ij}/k$. The overall mean is denoted $\bar{x}_.$ with $\bar{x}_. = \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij}/km$. The parameter α is estimated by $(\bar{x}_.)^{-1}$.

Data $\mathbf{y_1} = \{y_{1j}, j = 1, \ldots, n_c\}$ of n_c observations from one group from a crime scene (control data) and data $\mathbf{y_2} = \{y_{2j}, 1, j = \ldots, n_s\}$ of n_s observations from a group associated with a suspect (recovered data) are obtained. The value, V, of the evidence of these data is to be determined.

The exponential distribution is investigated as it is not easy to transform to a normal distribution and because a theoretical value for the likelihood ratio may be obtained against which various estimative procedures may be compared.

Let $\bar{y}_1 = \sum_{j=1}^{n_c} y_{1j}/n_c$ and $\bar{y}_2 \sum_{j=1}^{n_s} y_{2j}/n_s$ denote the means of the control and recovered data, respectively. Let $s_{y1}^2 = \sum_{j=1}^{n_c}(y_{1j} - \bar{y}_1)^2/(n_c - 1)$ and $s_{y2}^2 \sum_{j=1}^{n_s}(y_{2j} - \bar{y}_2)^2/(n_s - 1)$ denote the variances of the control and recovered data, respectively.

The within-group variance σ^2 of the underlying population is assumed known. Its value is taken to be $s_w^2 = \sum_{i=1}^{m} \sum_{j=1}^{k}(x_{ij} - \bar{x}_i)^2/(mk - m)$. The between-group variance of the underlying population is also assumed known. Its value is taken to be $s_b^2 = \sum_{i=1}^{m}(\bar{x}_i - \bar{x}_.)^2/(m-1) - s_w^2/k$.

The value of the evidence $(\mathbf{y_1}, \mathbf{y_2})$ is given by

$$V = \frac{\int f(\mathbf{y_1}, \mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu}{\int f(\mathbf{y_1} \mid \mu, \sigma 2) f(\mu \mid \alpha) d\mu \int f(\mathbf{y_2} \mid \mu, \sigma 2) f(\mu \mid \alpha) d\mu}$$

First, consider the term in the denominator for the control data $\mathbf{y_1}$; denote this term D_1. The within-group variance σ^2 is assumed known and the within-group distribution is assumed normal, thus the information in the data is contained in the sufficient statistic $\bar{Y}_1$. Then

$$D_1 = \int f(\bar{y}_1 \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu$$

where

$$f(\bar{y}_1 \mid \mu, \sigma^2) \frac{\sqrt{n_c}}{\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{n_c}{2\sigma^2}(\bar{y}_1 - \mu)^2 \Big\}.$$

Then

$$\begin{aligned} D_1 &= \frac{\alpha\sqrt{n_c}}{\sigma\sqrt{2\pi}} \int \exp\{-\frac{n_c}{2\sigma^2}(\bar{y}_1 - \mu)^2 - \alpha\mu\} d\mu \\ &= \frac{\alpha\sqrt{n_c}}{\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{\alpha}{2}(2\bar{y}_1 - \frac{\alpha\sigma^2}{n_c}) \Big\} \int \exp\Big[-\frac{n_c}{2\sigma^2} \Big\{ \mu - (\bar{y}_1 - \frac{\alpha\sigma^2}{n_c}) \Big\}^2 \Big] d\mu \\ &= \alpha \exp\Big\{ -\frac{\alpha}{2}(2\bar{y}_1 - \frac{\alpha\sigma^2}{n_c}) \Big\}. \end{aligned}$$

Similarly, the second term, denoted D_2, in the denominator, is given by

$$D_2 = \alpha \exp\Big\{ -\frac{\alpha}{2}(2\bar{y}_2 - \frac{\alpha\sigma^2}{n_s}) \Big\}.$$

Before considering the numerator, some extra notation is helpful.

$$\sigma^2_{12} = \sigma^2\Big(\frac{1}{n_c} + \frac{1}{n_s}\Big);$$

$$\begin{aligned} \sigma_3^2 &= \frac{\sigma^2}{n_c + n_s} + \frac{1}{\alpha^2}; \\ w &= (n_c\bar{y}_1 + n_s\bar{y}_2)/(n_c + n_s). \end{aligned}$$

If the between-group distribution of the data is not assumed to be exponential, the term $1/\alpha^2$ in the expression for σ_3^2 is replaced by the variance of the biweight kernel function. The derivation of this variance is described in Appendix 1.

The numerator, N, is $\int f(\mathbf{y_1}, \mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu$.

When $\mathbf{y_1}$ and $\mathbf{y_2}$ come from the same source, as is assumed in the numerator, they are dependent within the marginal distribution. As before, with σ^2 known, the information in the data is contained in the sufficient statistics $\bar{y}_1$ and $\bar{y}_2$. Following the argument of Lindley (1977), transform $\mathbf{y_1}, \mathbf{y_2}$ to independent statistics $(\bar{y}_1 - \bar{y}_2, w)$, with unit Jacobian. Also,

$$\begin{aligned} E(\bar{Y}_1 - \bar{Y}_2) &= 0; \\ Var(\bar{Y}_1 - \bar{Y}_2) &= \sigma^2\Big(\frac{1}{n_c} + \frac{1}{n_s}\Big) = \sigma_{12}^2; \\ E(W) &= \alpha^{-1}; \\ Var(W) &= \alpha^{-2} + \sigma^2\Big(\frac{1}{n_c + n_s}\Big) = \sigma_3^2. \end{aligned}$$

Thus

$$\begin{aligned} N &= \int f(\bar{y}_1 - \bar{y}_2) f(w \mid \mu) f(\mu \mid \alpha) d\mu \\ &= f(\bar{y}_1 - \bar{y}_2) \int f(w \mid \mu) f(\mu \mid \alpha) d\mu \\ &= f(\bar{y}_1 - \bar{y}_2) \frac{\alpha}{\sigma_3\sqrt{2\pi}} \int \exp\Big\{ -\frac{1}{2\sigma_3^2}(w - \mu)^2 - \alpha\mu \Big\} d\mu \end{aligned}$$

$$
\begin{aligned}
&= f(\bar{y}_1 - \bar{y}_2)\ \ \alpha\ \ \exp\Big\{\frac{\alpha}{2}(2w + \alpha\sigma_3^2)\Big\} \\
&= \frac{\alpha}{\sigma_{12}\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2\sigma_{12}^2}(\bar{y}_1 - \bar{y}_2)^2 + \frac{\alpha}{2}(2w + \alpha\sigma_3^2)\Big\}.
\end{aligned}
$$

The ratio $N/(D_1 D_2)$ gives the value, V, of the evidence as

$$
V = \frac{1}{\alpha\ \sigma_{12}\ \sqrt{2\pi}} \exp\Big[-\frac{1}{2\sigma_{12}^2}\big(\bar{y}_1 - \bar{y}_2\big)^2 + \frac{\alpha}{2}\Big\{2(w + \bar{y}_1 + \bar{y}_2) + \alpha\sigma_3^2 - \alpha\sigma^2\big(\frac{1}{n_c} + \frac{1}{n_s}\big)\Big\}\Big]. \tag{2}
$$

In what follows, the performance of the estimates of the value of the evidence obtained from various procedures will be compared with the theoretical value obtained from (2). For simulations, the theoretical value is determined with known α and σ. For estimations, using the techniques described below, the parameters α and σ^2 are replaced by their estimates from the population data $\{x_{ij}, i = 1, \ldots, m; j = 1, \ldots, k\}$, namely, $(\bar{x})^{-1}$ and s_w^2, respectively.

3 Estimation of likelihood ratio

If the between-group distribution is assumed to be exponential then an estimate of the value of evidence in a particular case with control data $\mathbf{y}_1$ and recovered data $\mathbf{y}_2$ may be obtained with substitution of the appropriate numerical values for $\bar{y}_1$ and $\bar{y}_2$ in (2).

In practice, a general approach is required which may then be applied to data which are highly positively skewed. Four different models for the between-group distribution are considered. Their values for the likelihood ratios are compared with the theoretical likelihood ratio. The within-group distribution

is considered to be normal throughout. The between-group distribution is taken to be one of

(i) a normal distribution, mean θ, variance τ^2 and

(ii) an exponential distribution with expectation estimated from population data.

3.1 Biweight kernel estimation

The use of a kernel density estimate based on the normal distribution is difficult when there is an achievable lower bound to the range of the variable being modelled and the data are highly positively skewed so that many of the data are close to the lower bound. In the example to be discussed here, the lower bound is zero and a kernel based on a normal distribution is very inaccurate close to this lower bound. A more appropriate approach for modelling a highly positively skewed distribution is the use of a biweight kernel (Wand and Jones, 1995) with a boundary kernel for use when the kernel comes close to the lower bound of the range of the random variable, in this case zero.

The biweight kernel $K(z)$ is defined as

$$K(z) = \frac{15}{16}(1 - z^2)^2; \quad | z | < 1. \tag{3}$$

This kernel is used to model the between-group distribution using the sample means $\{\bar{x}_1, \ldots, \bar{x}_m\}$. A general biweight kernel, with smoothing parameter h, and with a between-group variance of τ^2 is given by

$$\frac{1}{h\tau} K\Big(\frac{\mu - \bar{x}}{h\tau}\Big) = \frac{15}{16h\tau}\Big\{1 - \Big(\frac{\mu - \bar{x}}{h\tau}\Big)^2\Big\}^2; \quad \bar{x} - h\tau < \mu < \bar{x} + h\tau. \tag{4}$$

There are two candidates for the estimation of the between-group variance,

(i) $s_b^2 = \sum_{i=1}^{m}(\bar{x}_i - \bar{x}_.)^2/(m-1) - s_w^2/k$,

(ii) $1/(\bar{x})^2$,

the least-squares estimate and the method of moments estimate, respectively, of τ^2, the between-group variance.

The problem of a fixed lower bound at zero is tackled with a boundary kernel. When an observation, $\bar{x}$, is close to zero, a different kernel, known as the boundary kernel (Wand and Jones, 1995), is used. Closeness is defined as $\bar{x} < h\tau$. For $\bar{x} > h\tau$, the biweight kernel (4) is used. For $\bar{x} < h\tau$, a boundary kernel

$$K_h(z) = \frac{\nu_2 - \nu_1 z}{\nu_0\nu_2 - \nu_1^2} K(z) \tag{5}$$

is used where $K(z)$ is as given in (3). For ease of notation, denote $h\tau$ by δ. The terms ν_0, ν_1 and ν_2 are constants, functions of δ. For the kernel (3) these are defined as

$$\nu_t = \int_{-1}^{\delta} z^t K(z) dz, \;\; t = 0, 1, 2,$$

where the dependency of ν on δ is suppressed. They can be shown to be

$$\begin{aligned}
\nu_2 &= \frac{1}{14}\big\{1 + \frac{1}{8}\delta^3(35 - 42\delta^2 + 15\delta^4)\big\}, \\
\nu_1 &= \frac{5}{32}\big\{\delta^2(3 - 3\delta^2 + \delta^4) - 1\big\}, \\
\nu_0 &= \frac{1}{2} + \frac{15}{16}(\delta - \frac{2}{3}\delta^3 + \frac{1}{5}\delta^5).
\end{aligned}$$

In practice, the factor $(\nu_2 - \nu_1 z)/(\nu_0\nu_2 - \nu_1^2)$ is close to 1.

An optimal value of the smoothing parameter h is given by

$$h_{opt} = \Big(\frac{1}{7}\Big)^{-\frac{2}{5}}\Big(\frac{15}{21}\Big)^{\frac{1}{5}}\Big\{\int f''(x)^2 dx\Big\}^{-\frac{1}{5}} m^{-\frac{1}{5}}$$

(Silverman, 1986). Then, it can be shown that, when $f(x)\alpha \exp\{-\alpha x\}$,

$$h_{opt} = \Big(\frac{70}{m}\Big)^{\frac{1}{5}}\alpha^{-1}$$

which can be estimated by

$$h_{opt} = \Big(\frac{70}{m}\Big)^{\frac{1}{5}}\bar{x}.$$

.

3.2 Likelihood ratio with biweight and boundary kernels

3.2.1 Biweight kernel

First, consider the denominator and the factor which is associated with the control sample $\{y_{1i}, i1, \ldots, n_c\}$. Denote this as D_c. This may be written as

$$D_c = \int f(y_{11}, \ldots, y_{1n_c} \mid \mu,\, \sigma^2) f(\mu \mid \alpha) d\mu.$$

The factor associated with the recovered sample may be derived analogously and denote this as D_s. The between-group exponential distribution $f(\mu \mid \alpha)$ is replaced with the kernel

$$\hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m) = \frac{1}{mh\tau}\sum_{i=1}^{m} K\Big(\frac{\mu - \bar{x}_i}{h\tau}\Big) = \frac{1}{mh\tau}\sum_{i=1}^{m}\Big\{1 - \Big(\frac{\mu - \bar{x}_i}{h\tau}\Big)^2\Big\}^2. \quad (6)$$

It is convenient to make a transformation $z_i = (\mu - \bar{x}_i)/(h\tau)$ with $\mu = \bar{x}_i + h\tau z_i$, Jacobian $d\mu = h\,\tau\,dz_i$, and $-1 < z_i < 1$.

The distribution of Y, for both control and recovered sources, conditional on μ, is normal so as for the derivation of (2), only the distribution of the sufficient statistic $\bar{Y}$ need be considered for the distributions of the terms in the expression for the likelihood ratio.

The first term, D_c, in the denominator, with the biweight kernel (4) used for $f(\mu \mid \alpha)$, is given by

$$\begin{aligned} D_c &= \int f(\bar{y}_1 \mid \mu,\ \sigma^2)\hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m))d\mu \\ &= \frac{\sqrt{n_c}}{\sigma\sqrt{2\pi}} \int \exp\Big\{ -\frac{n_c}{2\sigma^2}(\bar{y}_1 - \mu)^2 \Big\} \Big[\frac{15}{16\ m\ h\ \tau} \sum_{i=1}^{m} \Big\{ 1 - \Big(\frac{\mu - \bar{x}_i}{h\ \tau} \Big)^2 \Big\}^2 \Big] d\mu \\ &= \frac{15\ \sqrt{n_c}}{16\ m\ \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} (1 - z_i^2)^2 \exp\Big\{ -\frac{n_c}{2\sigma^2}(\bar{y}_1 - (\bar{x}_i + h\ \tau\ z_i))^2 \Big\} d\ z_i. \end{aligned}$$

Similarly, the second term, D_s, in the denominator, with the biweight kernel (4) used for $f(\mu \mid \alpha)$, is given by

$$\begin{aligned} D_s &= \int f(\bar{y}_2 \mid \mu,\ \sigma^2)\hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m)d\mu \\ &= \frac{15\ \sqrt{n_s}}{16\ m\ \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} (1 - z_i^2)^2 \exp\Big\{ -\frac{n_s}{2\sigma^2}(\bar{y}_2 - (\bar{x}_i + h\ \tau\ z_i))^2 \Big\} d\ z_i. \end{aligned}$$

Now, consider the numerator. Denote this as N_{cs}. As previously

$$N_{cs} = f(\bar{y}_1 - \bar{y}_2) \int f(w \mid \mu)\hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m)d\mu.$$

Then

$$\int f(w \mid \mu)\hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m)d\mu =$$

$$\frac{15}{16\ m\ h\ \tau}\frac{1}{\sigma_3\sqrt{2\pi}}\sum_{i=1}^{m}\int \exp\Big\{-\frac{1}{2\sigma_3^2}(w-\mu)^2\Big\}\Big\{1-\big(\frac{\mu-\bar{x}_i}{h\ \tau}\big)^2\Big\}^2 d\mu$$

$$= \frac{15}{16\ m\ \sigma_3\sqrt{2\pi}}\sum_{i=1}^{m}\int_{-1}^{1}(1-z_i^2)^2\exp\Big[-\frac{1}{2\sigma_3^2}\big\{w-(\bar{x}_i+h\ \tau z_i)\big\}^2\Big]d\ z_i.$$

Thus

$$N_{cs} = \frac{1}{\sigma_{12}\ \sqrt{2\pi}}\exp\Big\{-\frac{1}{2\sigma_{12}^2}(\bar{y}_1-\bar{y}_2)^2\Big\}\frac{15}{16\ m\ \sigma_3\ \sqrt{2\pi}}$$
$$\sum_{i=1}^{m}\int_{-1}^{1}(1-z_i^2)^2\exp\Big[-\frac{1}{2\sigma_3^2}\big\{w-(\bar{x}_i+h\ \tau\ z_i)\big\}^2\Big]dz_i.$$

The likelihood ratio is then given by the ratio of N_{cs} to the product of D_c and D_s. Numerical evaluation of the likelihood ratio may then be made with the substitution of σ by s_w, τ by s_b and h by its optimal value $(70/m)^{1/5}\bar{x}$.

3.2.2 Boundary kernel

There is a boundary effect when $(\bar{x}_i, i = 1, \ldots, m)$ is within $h\tau$ of zero. For such $\bar{x}_i$, the kernel expression

$$\Big\{1-\Big(\frac{\mu-\bar{x}_i}{h\tau}\Big)^2\Big\}^2 = \Big\{1-z_i^2\Big\}^2$$

has to be adjusted with the factor $(\nu_2-\nu_1 z)/(\nu_0\nu_2-\nu_1^2)$, where $z_i = (\mu-\bar{x}_i)/(h\tau)$ and ν_0, ν_1, ν_2 are as in (5), to give

$$\frac{(\nu_2-\nu_1 z_i)}{(\nu_0\nu_2-\nu_1^2)}\Big\{1-z_i^2\Big\}^2$$

which can be written as

$$(a-bz_i)\Big\{1-z_i^2\Big\}^2$$

where $a = \nu_2/(\nu_0\nu_2 - \nu_1^2)$ and $b\nu_1/(\nu_0\nu_2 - \nu_1^2)$. Define an indicator function $\gamma(z_i)$ such that

$$\begin{aligned}\gamma(z_i) &= 1 \text{ if } x_i > h\tau, \\ &= (a - bz_i) \text{ if } x_i < h\tau.\end{aligned}$$

Then the likelihood ratio $N_{cs}/(D_cD_s)$ can be adapted to account for boundary effects to give a value for the evidence of

$$\frac{1}{\sigma_{12}\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2\sigma_{12}^2}(\bar{y}_1 - \bar{y}_2)^2 \Big\} \frac{15}{16\, m\, \sigma_3\, \sqrt{2\pi}}$$
$$\sum_{i=1}^{m} \int_{-1}^{1} \gamma(z_i)(1 - z_i^2)^2 \exp\Big[-\frac{1}{2\sigma_3^2}\big\{w - (\bar{x}_i + h\,\tau\, z_i)\big\}^2 \Big] dz_i,$$

divided by the product of

$$\frac{15\,\sqrt{n_c}}{16\, m\, \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} \gamma(z_i)(1 - z_i^2)^2 \exp\Big\{ -\frac{n_c}{2\sigma^2}(\bar{y}_1 - (\bar{x}_i + h\,\tau\, z_i))^2 \Big\} d\, z_i$$

and

$$\frac{15\,\sqrt{n_s}}{16\, m\, \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} \gamma(z_i)(1 - z_i^2)^2 \exp\Big\{ -\frac{n_s}{2\sigma^2}(\bar{y}_2 - (\bar{x}_i + h\,\tau\, z_i))^2 \Big\} d\, z_i.$$

4 Simulations

The likelihood ratio is calculated for various different scenarios. The between-group distribution for μ is assumed to be exponential with density function $f(\mu \mid \alpha) = \alpha \exp(-\alpha\mu)$. The within-group distribution is assumed to be normal

with expectation μ and variance σ^2. Four different scenarios investigated are as follows.

- Control data $\{y_{1i}, i = 1, \ldots, n_c\}$ and recovered data $\{y_{2i}, i = 1, \ldots, n_s\}$ are generated from around the mean μ of one of the groups and μ is generated from close to the population mean $1/\alpha$.

- Control data $\{y_{1i}, i = 1, \ldots, n_c\}$ and recovered data $\{y_{2i}, i = 1, \ldots, n_s\}$ are generated from the periphery of one of the groups (*e.g.*, outside the 95 percentile of the associated normal distribution) and μ is generated from close to the population mean $1/\alpha$.

- Control data $\{y_{1i}, i = 1, \ldots, n_c\}$ and recovered data $\{y_{2i}, i = 1, \ldots, n_s\}$ are generated from around the mean μ of one of the groups and μ is generated from the periphery of the population (*e.g.*, outside the 95 percentile of the exponential distribution with expectation $1/\alpha$).

- Control data $\{y_{1i}, i = 1, \ldots, n_c\}$ and recovered data $\{y_{2i}, i = 1, \ldots, n_s\}$ are generated from the periphery of one of the groups (*e.g.*, outside the 95 percentile of the associated normal distribution) and μ is generated from the periphery of the population (*e.g.*, outside the 95 percentile of the exponential distribution with expectation $1/\alpha$).

In all these cases, the control and recovered data come from the same group. Thus, the likelihood ratios should all be greater than one.

Population data $\{x_{ij}, i = 1, \ldots, m, j = 1, \ldots, k\}$ are generated from a two-level model in which the between-group distribution for μ is taken to be exponential with density function $f(\mu \mid \alpha) = \alpha \exp(-\alpha\mu)$ and the within-group distribution is taken to be normal with mean μ, generated from the exponential distribution, and variance σ^2. Values for m, k and α are 50, 10 and 1.0 respectively, The former two are taken to be representative of values which may be seen in casework and the value of 1.0 for α is taken as a value which provides a reasonable level of skewness.

The likelihood ratios for the four generated control and recovered data scenarios are calculated for

(i) normal distribution, mean μ, variance τ^2;

(ii) exponential distribution with expectation estimated from the population data;

(iii) non-normal distribution, estimated by a normal kernel function as described in Aitken and Taroni (2004), adapted to allow for the correlation between the control and recovered data $\bar{y}_1$ and $\bar{y}_2$ if they are assumed, as in the numerator, to come from the same source and extended to an adaptive kernel (see Section 4.1);

(iv) non-normal distribution, estimated by a biweight kernel function with a boundary kernel as described in Silverman (1986) and Wand and Jones (1995), and detailed in Section 3.2.

4.1 Non-normal between-group distribution with normal adaptive kernel function

The value of the evidence, when the between-group distribution is taken to be non-normal and is estimated by a normal kernel function as described in Aitken and Taroni (2004, equation (10.12)), is adapted to allow for the correlation between the control and recovered data $\bar{y}_1$ and $\bar{y}_2$ if they are assumed, as in the numerator, to come from the same source. A multivariate version of this formulation is given in Aitken *et al.* (2007). This expression is then extended to an adaptive kernel, where the smoothing parameter is dependent on x_i and is thus denoted h_i.

The numerator is

$$\frac{1}{m}(2\pi)^{-1}\{(\frac{n_c+n_s}{n_c\ n_s})\sigma^2\}^{-1/2}\{\tau^2+\frac{\sigma^2}{n_c+n_s}\}^{-1/2}(h_i^2\tau^2)^{-1/2}$$
$$\{(\tau^2+\frac{\sigma^2}{n_c+n_s})^{-1}+(h_i^2\tau^2)^{-1}\}^{-1/2}\exp\{-\frac{1}{2}(\bar{y}_1-\bar{y}_2)^2[(\frac{n_c+n_s}{n_c\ n_s})\sigma^2]^{-1}\}$$
$$\sum_{i=1}^{m}\exp\{-\frac{1}{2}(w-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_c+n_s}+h_i^2\tau^2)^{-1}\}.$$

The first term in the denominator is

$$\frac{1}{m}(2\pi)^{-1/2}\{\tau^2+\frac{\sigma^2}{n_c}\}^{-1/2}(h_i^2\ \tau^2)^{-1/2}$$
$$\{(\tau^2+\frac{\sigma^2}{n_c})^{-1}+(h_i^2\tau^2)^{-1}\}^{-1/2}$$
$$\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\bar{y}_1-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_c}+h_i^2\tau^2)^{-1}\}.$$

The second term in the denominator is

$$\frac{1}{m}(2\pi)^{-1/2}\{\tau^2+\frac{\sigma^2}{n_s}\}^{-1/2}(h_i^2\ \tau^2)^{-1/2}$$
$$\{(\tau^2+\frac{\sigma^2}{n_s})^{-1}+(h_i^2\tau^2)^{-1}\}^{-1/2}$$
$$\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\bar{y}_2-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_s}+h_i^2\tau^2)^{-1}\}.$$

The constant term in the ratio is then:

$$\frac{m\Big\{n_c\tau^2(h_i^2+1)+\sigma^2\Big\}^{1/2}\Big\{n_s\tau^2(h_i^2+1)+\sigma^2\Big\}^{1/2}}{\sigma\Big\{(n_c+n_s)\tau^2(h_i^2+1)+\sigma^2\Big\}^{1/2}}.$$

The remaining term, that involving $\bar{y}_1, \bar{y}_2$ and $\bar{x}_i$, is the ratio of

$$\exp\{-\frac{1}{2}(\bar{y}_1-\bar{y}_2)^2(\sigma^2(\frac{1}{n_c}+\frac{1}{n_s}))^{-1}\}\sum_{i=1}^{m}\exp\{-\frac{1}{2}(w-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_c+n_s}+h_i^2\tau^2)^{-1}\}$$

to

$$\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\bar{y}_1-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_c}+h_i^2\tau^2)^{-1}\}\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\bar{y}_2-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_s}+h_i^2\tau^2)^{-1}\}.$$

The adaptive smoothing parameter h_i is estimated using the procedure outlined in Silverman (1986).

First, a pilot estimate $\tilde{f}(x)$ is obtained with a kernel density estimation procedure using a Gaussian kernel, which automatically satisfies the condition that $\tilde{f}(x_i) > 0$ for all i. The smoothing parameter h_i is then defined by

$$h_i = \left\{\tilde{f}(x_i)/g\right\}^{-\beta}$$

where g is the geometric mean of the $\tilde{f}(x_i)$:

$$\log g = m^{-1} \sum \log \tilde{f}(x_i)$$

and β is a sensitivity parameter, a number satisfying $0 \leq \beta \leq 1$.

These likelihood ratios are compared with the theoretical values (2) by taking the ratio of the estimated value to the theoretical value. Values for this ratio close to one are good, values less than one show that the estimated value is underestimating the true value, values greater than one show that the estimated value is overestimating the true value.

4.2 Results

Likelihood ratios are calculated for data in which the between-group distribution is exponential, with parameter α, and the within-group distribution is normal with variance σ^2 and whose mean is exponentially distributed. Control data are $y_{1j}, j = 1, \ldots, n_c$ where $n_c = 5$. Recovered data are $y_{2j}, j = 1, \ldots, n_s$ where $n_s = 5$. The overall population mean is α^{-1} and population variance is α^{-2}. The value 1.0 is used for α.

Nine pairs of control and recovered data are used. First, the control data are simulated from normal distributions with expectations $\alpha^{-1}, \alpha^{-1}+\alpha^{-1}$ and $\alpha^{-1}+2\alpha^{-1}$; *i.e.*, at the mean of the between-group distribution and then one between-group standard deviation and two between-group standard deviations away from the mean, and with variance σ^2. The recovered data are simulated from normal

distributions with expectations $\bar{y}_1$, $\bar{y}_1+\sigma$ and $\bar{y}_1+2\sigma$; *i.e.*, at the sample mean of the control data, and then one within-group standard deviation and two within-group standard deviations of that sample mean. All nine combinations of control and recovered data are simulated. There are 500 simulations of each combination in total. The purpose of the simulations is to illustrate the changes in the likelihood ratio for various combinations of similarity and rarity. Similarity is when the distribution of $\mathbf{y_2}$ has the mean $\bar{y}_1$ for example. Rarity is when the control data are simulated from a normal distribution with expectation $\alpha^{-1} + 2\alpha^{-1}$.

The theoretical likelihood ratio (2) for the nine combinations is determined as is the likelihood ratio using a biweight kernel, a normal kernel (Section 4.1) and an assumption of between group normality with the between-group expectation and variance taken to be α^{-1} and α^{-2}, respectively. An adaptive kernel was investigated with the adaptive parameter β taking values $0, 0.1, 0.2$ and 0.5. Results are given in Table 4.2

The adaptive kernel model of the between-group distribution provides reasonable results, as measured by the ratio of its estimated value to the true value, at the expectation of the exponential. Results at one or two standard deviations from the expectation are not so good. The normal model of the between-group distribution does not provide good results. The biweight kernel model gives the best results at two standard deviations from the expectation ($\alpha^{-1}+2\alpha^{-1}$), very good results at one standard deviation from the expectation ($\alpha^{-1} + \alpha^{-1}$) but not such good results as the adaptive kernel at the expectation.

Table 1: Means of 500 simulations of likelihood ratios for evidence from a two-level model in which the within-group distribution is distributed normally with variance σ^2 and with expectation which has a between-group distribution that is exponential with expectation α^{-1}. The control data are five simulations from normal distributions which have expectations $\alpha^{-1}, \alpha^{-1}+\alpha^{-1}$, and $\alpha^{-1}+2\alpha^{-1}$, and constant variance σ^2. The sample means of these simulations are denoted $\bar{y}_1$. The recovered data are five simulations from normal distributions which have expectations $\bar{y}_1$, $\bar{y}_1+\sigma$ and $\bar{y}_1+2\sigma$. The sample means of these simulations are denoted $\bar{y}_2$. The parameter for the adaptive kernel is denoted β and takes values 0 (corresponding to a standard kernel), 0.1, 0.2 and 0.5. The ratios of the estimated values to the corresponding theoretical values are given in parentheses alongside the appropriate estimated value.

Expectation of control data $\bar{y}_1$	Adaptive parameter	Expectation of recovered data $\bar{y}_2$		
		$\bar{y}_1$	$\bar{y}_1+\sigma$	$\bar{y}_1+2\sigma$
		Between group modelled by adaptive kernel		
	β			
α^{-1}	0	10.19 (0.79)	3.37 (0.76)	0.15 (0.75)
	0.1	10.17 (0.79)	3.37 (0.76)	0.15 (0.75)
	0.2	10.19 (0.79)	3.38 (0.76)	0.15 (0.75)
	0.5	10.51 (0.81)	3.50 (0.79)	0.16 (0.80)
$\alpha^{-1}+\alpha^{-1}$	0	20.14 (0.43)	8.92 (0.43)	0.04 (0.40)
	0.1	20.45 (0.44)	9.06 (0.44)	0.04 (0.40)
	0.2	20.81 (0.45)	9.23 (0.45)	0.04 (0.40)
	0.5	22.70 (0.49)	10.09 (0.49)	0.05 (0.50)
$\alpha^{-1}+2\alpha^{-1}$	0	50.48 (0.35)	13.32 (0.35)	0.06 (0.38)
	0.1	51.10 (0.36)	13.47 (0.36)	0.06 (0.38)
	0.2	51.83 (0.36)	13.65 (0.36)	0.06 (0.38)
	0.5	56.49 (0.40)	14.83 (0.39)	0.06 (0.38)
		Between and within both normal		
α^{-1}		72.82 (5.64)	23.96 (5.40)	1.05 (5.25)
$\alpha^{-1}+\alpha^{-1}$		156.29 (3.40)	70.39 (3.41)	0.34 (3.40)
$\alpha^{-1}+2\alpha^{-1}$		791.85 (5.56)	226.48 (5.97)	0.99 (6.19)
		Between group modelled by biweight kernel		
α^{-1}		5.21 (0.40)	1.84 (0.41)	0.09 (0.45)
$\alpha^{-1}+\alpha^{-1}$		37.25 (0.81)	16.92 (0.82)	0.09 (0.90)
$\alpha^{-1}+2\alpha^{-1}$		244.27 (1.72)	59.47 (1.57)	0.24 (1.50)
		Theoretical value		
α^{-1}		12.92	4.44	0.20
$\alpha^{-1}+\alpha^{-1}$		46.01	20.66	0.10
$\alpha^{-1}+2\alpha^{-1}$		142.40	37.91	0.16

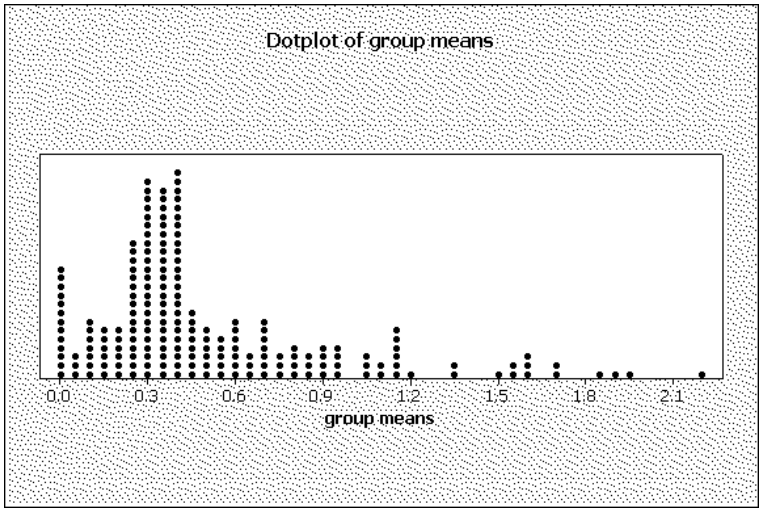

Figure 1: Dotplot of group means

5 Application

Figure 1 shows the distribution of the concentrations of aluminium in 200 groups of glass fragments with 12 fragments in each group; hence in our notation, $m = 200, k = 12$. The data plotted are the 200 group means. Whilst the data are not truly exponential since the mode is slightly removed from zero, the data are highly positively skewed.

The performance of three procedures for estimating the likelihood ratio is compared. The three procedures estimate the between-group distributions with a normal distribution, a normal adaptive kernel and a biweight kernel with a boundary condition. There is, of course, no theoretical model with which to compare the results and assess their performance. Instead, control and recovered data are taken from the overall sample data as follows: for same-source

comparisons, the control and recovered data are taken from the same group by splitting the group into two equally-sized, non-overlapping halves (containing two measurements each); for different-source comparisons, the control and recovered data are entire groups selected from different sources.

There are 200 within-group comparisons of control and recovered data and $200 \times 199/2 = 19,900$ between-group comparisons. For the 200 within-group comparisons, the likelihood ratio should be greater than 1 and for the 19,900 between-group comparisons, the likelihood ratio should be less than 1. The results are shown in Table 5.

The biweight kernel has the largest false negative rate (91/200,45.5%) and the lowest false positive rate (3103/19,900, 15.5%) of the various models. In a criminal trial, it is more important to have a small false positive rate (wrongful conviction of an innocent person) than a small false negative rate (wrongful release of a guilty person).

6 Conclusions

At present, likelihood ratios for two-level models are determined with the use of a normal kernel estimation procedure when the between-group distribution is thought to be non-normal. An extension is described here for a two-level model in which the between-group distribution is very positively skewed and an exponential distribution may be thought to represent a good model. A biweight kernel model is shown to provide results which are better than a normal kernel model and comparable to an adaptive kernel model.

Table 2: Summary of likelihood ratios for aluminium data. Two hundred calculations of within-group comparisons and 19,900 calculations of between-group comparisons are made. Results are recorded for a normal kernel estimation (nn), an exponential kernel estimation (exp), an adaptive kernel estimation for $\beta = 0, 0.1, 0.2$ and 0.5 and a biweight kernel estimation (b).

Likelihood ratio range	nn	exp	β				b
			0.0	0.1	0.2	0.5	
			Within-group comparisons				
$0-1$	4	4	4	4	4	4	91
$1-10^1$	1	160	184	184	184	185	76
10^1-10^2	183	35	12	12	12	11	32
10^2-10^3	8	1	0	0	0	0	1
10^3-10^4	3	0	0	0	0	0	0
$>10^4$	1	0	0	0	0	0	0
			Between-group comparisons				
$<10^{-4}$	6082	6479	6515	6514	6513	6499	6672
$10^{-4}-10^{-3}$	830	1200	1129	1129	1143	1123	1414
$10^{-3}-10^{-2}$	1282	1722	1716	1716	1706	1642	2171
$10^{-2}-10^{-1}$	1856	2384	2518	2511	2501	2476	5651
$10^{-1}-1$	619	763	809	809	796	789	889
$1-10^1$	2972	7031	7153	7162	7183	7306	2834
$>10^1$	6259	321	60	59	58	65	269

7 Acknowledgements

The aluminium data were provided by Dr. G. Zadora of the Forensic Research Institute in Cracow, Poland, and we are grateful to the Institute and to him for permission to use the data and for helpful discussions. The work has been supported by the EPSRC as part of their programme on Technologies for Crime Prevention and Detection, grant reference GR/S98603.

8 Appendix 1

8.1 The derivation of the variance of the biweight kernel

The between-group exponential distribution is replaced with the kernel:

$$\hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m)\frac{1}{mh\tau}\sum_{i=1}^{m} K\Big(\frac{\mu - \bar{x}_i}{h\tau}\Big) = \frac{1}{mh\tau}\sum_{i=1}^{m}\Big\{1 - \Big(\frac{\mu - \bar{x}_i}{h\tau}\Big)^2\Big\}^2.$$

The variance of this distribution is $E(\mu^2) - \{E(\mu)\}^2$, where

$$\begin{aligned} E(\mu^k) &= \int \mu^k \hat{f}(\mu \mid \bar{x}_1, \ldots, \bar{x}_m)d\mu; \quad k = 1, 2; \\ &= \frac{1}{mh\tau}\sum_{i=1}^{m}\int_{\bar{x}_i - h\tau}^{\bar{x}_i + h\tau} \mu^k\Big\{1 - \Big(\frac{\mu - \bar{x}_i}{h\tau}\Big)^2\Big\}^2 d\mu. \end{aligned}$$

9 References

Aitken,C.G.G. and Taroni,F. (2004) *Statistics and the evaluation of evidence for forensic scientists*, 2nd edition, John Wiley and Sons, Ltd., Chichester.

Aitken,C.G.G., Zadora,G. and Lucy,D. (2007) A two-level model for evidence evaluation. *Journal of Forensic Sciences* (to appear.)

Lindley,D.V. (1977) A problem in forensic science. *Biometrika*, **64**, 207-213.

Silverman,B.W. (1986) *Density estimation*, Chapman & Hall, London.

Wand,M.P. and Jones,M.C. (1995) *Kernel Smoothing*, Chapman and Hall, London.